Teaching Through Tagging — Interactive Lexical Semantics

Francis Bond, Andrew Kirkrose Devadason,
Teo Rui Lin, Melissa and **Luís Morgado da Costa**
School of Humanities, Nanyang Technological University
bond@ieee.org,andrewtemerarious@gmail.com
trlm31@outlook.com,lmorgado.dacosta@gmail.com

Abstract

In this paper we discuss an ongoing effort to enrich students' learning by involving them in sense tagging. The main goal is to lead students to discover how we can represent meaning and where the limits of our current theories lie. A subsidiary goal is to create sense tagged corpora and an accompanying linked lexicon (in our case wordnets). We present the results of tagging several texts and suggest some ways in which the tagging process could be improved. Two authors of this paper present their own experience as students. Overall, students reported that they found the tagging an enriching experience. The annotated corpora and changes to the wordnet are made available through the NTU multilingual corpus and associated wordnets (NTU-MC).

1 Introduction

This paper introduces a method of incorporating lexical semantic research into the teaching of semantics, as a form of experiential learning (Kolb, 1984). The main goal is to lead students to discover how we can represent meaning and where the limits of our current theories lie. A subsidiary goal is to create sense tagged corpora and an accompanying linked lexicon (in our case wordnets).

The first author (Francis) teaches HG2002: Semantics and Pragmatics, a core course in linguistics with 70-100 students. The course is survey-oriented (Pullum, 1984, p152), summarising various theories without dwelling on any overarching theme. It is easy for students to become bewildered by the variety of concepts, particularly in the absence of concrete applications. To alleviate this, from 2011, each semester a text was introduced, which students would try to analyse using the various approaches. The decision was also influenced by our university's encouraging stance towards involving students in research. The NTU computational linguistics lab is heavily involved in lexical semantics and wordnets, including building and extending wordnets and sense tagged corpora for multiple languages. We thus integrated some tagging into the course as a way to give students hands-on experience with a semantics-oriented research project.

This course later formed the base for a general elective (GE), which is offered to any student in the university.[1] This was an interdisciplinary course, developed and co-taught with colleagues from the English and Chinese departments. To make it more appealing, we focused on Sherlock Holmes — HG8011: Detecting Meaning with Sherlock Holmes. Roughly half the course deals with interpreting the texts using semantics, a quarter with placing the stories in their literary context, including a discussion of fan-fiction, and the rest with Sherlock in film and in translation. The course has proved popular, with over 200 students every time it is offered, and long wait lists.

The pedagogical goals for both courses were fourfold:

P-1 Apply semantic theories to real-world texts

P-2 Show students the difficulties of defining and identifying senses. For example they need to look at more than prototypical cases; identify gaps in the lexicons and add new entries; and consider the problems of tokenization and MWEs.

P-3 Expose them to annotation and resource building (common sources of employment for humanities students)

P-4 Teach the students about inter-annotator agreement

There were four main research goals:

R-1 Produce sense tagged corpora, with all concepts disambiguated, in multiple languages

R-2 Experiment with how to sense tag: What is the best interface? What information do annotators need?

[1] Due to the content overlap with *Semantics and Pragmatics* a student cannot take both.

R-3 Identify interesting phenomena that can lead to student assignments or theses

R-4 Identify potential student research assistants

For teaching a subject like this, it is impossible to do a quantitative evaluation where half the class does the annotation and half does not. Instead, in this paper two students who took these classes share their experiences as students in Section 3. They both did well in the subjects and are keen on working further with wordnets. As such, their expressed views may not be representative of the student population at large. Therefore, we also looked at comments by the students in their assignments, and in the anonymous student course evaluation.

One project that is very similar to our annotation in spirit is the Georgetown University Multilayer Corpus (GUM: Zeldes, 2017). GUM is collected and expanded by students as part of the curriculum in LING-367 Computational Corpus Linguistics at Georgetown University. The course has around 20 students, mainly postgraduate. The students are more computational than in our courses, so there is more emphasis on using external tools to annotate. The corpus selection is opportunistic, aiming to represent different communicative purposes, while coming from sources that are readily and openly available (mostly Creative Commons licenses). The results of this project show that high quality, richly annotated resources can be created effectively as part of a linguistics curriculum. The main difference is that the course at Georgetown is specifically about corpora, while for my courses, the corpora are not the main focus.

Wordnets have also been used widely in teaching computational linguistics (Lemnitzer and Kunze, 2004; Bird et al., 2009, 2010), but as far as we know this is the first time they have been a core part of a general linguistics course.

The paper is structured as follows: in Section 2 we describe the actual practice of annotation. In Section 3 we present the student experience. In Section 4 we look at what work is necessary to make the corpus ready for release. We finish with some conclusions and ideas for future work in Section 5.

2 Annotating Texts

There is some overlap between the linguistics and GE course, but enough differences that we will describe them separately. Our university teaches in English, and the majority of the students are native speakers of English, although we have some international students (more in the GE class). Many students are also fluent in another mother tongue (mainly Mandarin Chinese, Standard Malay and sometimes Tamil) and many of the linguistics students have studied a second language to a level in which they can annotate meaning (with Japanese and Korean being the most popular).

2.1 Linguistic Students

Each year, students read one (or part) of a given texts After reading the text, and hearing lectures on word and sentence level semantics, each student tags a short passage (roughly 300 concepts: 20-30 sentences). Most years we choose a text that can be completely tagged by the class, so typically 600-700 sentences. The students found the specialist computer science content in the Cathedral and the Bazaar hard to understand, and much preferred either locally salient text (like the Singapore Tourist Data) or short stories. In 2015 we had a very multilingual group so we picked a shorter story and the students annotated the original Japanese as well as Chinese, English, and Malay translations. From 2018 we tagged a longer novel.

The texts annotated are listed below:

2011 Singapore Tourist Data (website)

2012 The Cathedral and the Bazaar (essay) (Raymond, 1999)

2013 The Adventure of the Speckled Band (Doyle, 1892)

2014 The Adventure of the Dancing Men (Doyle, 1905)

2015 蜘蛛の糸 *Kumo no Ito* “The Spider’s Web” (芥川, 1918)

2018–2020 The Hound of the Baskervilles (Doyle, 1902)

Each sentence is assigned at least three annotators. At first, three or four students tagged each passage. Since 2018 we have added an automatic tagger as the third annotator. This gives the students experience with automatic sense disambiguation, and allows us to tag more text. In order to make the automatic tagging predictable, we used a simple most-frequent sense based annotator, trained on frequencies in Princeton Wordnet combined with the already tagged short stories.

During the tagging, students look at every content word and find its corresponding meaning in a dictionary (wordnet). If there is an appropriate sense, then they select it. When such a meaning is absent from the wordnet, a new synset should be proposed. For the last three years, students have also annotated positive or negative sentiment (-100 to $+100$) at the sense level, using the set up described in Bond et al. (2016a). If there is an error in the corpus (such as incorrect tokenization or lemmatization or just a typo) the student tag it as '**e**', if there is a problem with the wordnet (no appropriate sense or indistinguishable senses) the students tag it as '**w**'. If a word should not be tagged (for example if it is a closed class word such as preposition or auxiliary) then it is tagged as '**x**'.

When students complete tagging individually, we calculate and show the agreement. A new text is made tagged with the majority tag for each concept, and students must then retag anything with no majority tag (and can, of course, retag anything at all). If any two taggers agree, their tag is selected: the automatic tagger thus only has an effect when two students disagree. Students tagging the same sentences meet up to discuss disagreements and then retag. Overall, the tagging takes roughly 5-6 hours for each round.

Finally, they write up a joint report on their findings (worth 30% of their final grade). In the final write-up, the students are asked to: (i) describe the strengths and weaknesses of using a lexical resource such as wordnet to define word meaning, (ii) give concrete examples from the text you analyzed. (iii) discuss cases where you disagreed with other annotators, on reflection, do you think: you were right; they were right; the definition is bad; or is there some other reason? (iv) For words with senses missing in wordnet, they should write a comment with enough information to create a new entry for them consisting of, at minimum, a definition, a relational link to an existing synset and an example.

2.2 General Elective Students

The GE students have no tutorials, and generally are expected to cover the material at a slightly easier level. For this class, students only tag Sherlock Holmes stories, and only in English.

NTU offers elective classes as part of General Education with discipline branches in Liberal Arts, Science and Technology, and Business. *HG8011: Detecting Meaning with Sherlock Holmes* falls under Liberal Arts. The course teaches semantics, some literature, film theory and translation studies. The assignments follow the same structure as *HG2002: Semantics and Pragmatics*, except that a written report is not required. The stories tagged are:

2016 The Redheaded League
(Doyle, 1892)

2018 A Scandal in Bohemia
(Doyle, 1892)

2019 The Hound of the Baskervilles
(Doyle, 1902)

The stories are chosen from the most popular of the short stories (Doyle, 1927), plus the most popular novel.

The project is broken into three parts for these students: tag individually (20%), tag as a group (20%), tag sentiment (20%). Each passage is given to three or four students as the drop out rate for general electives is around 10% — this means that some groups end up with fewer than three for the comparison. We also add the automatic tagger. The GE students are not asked to write a report, instead they are judged on the comments they enter when they tag.

2.3 Interface

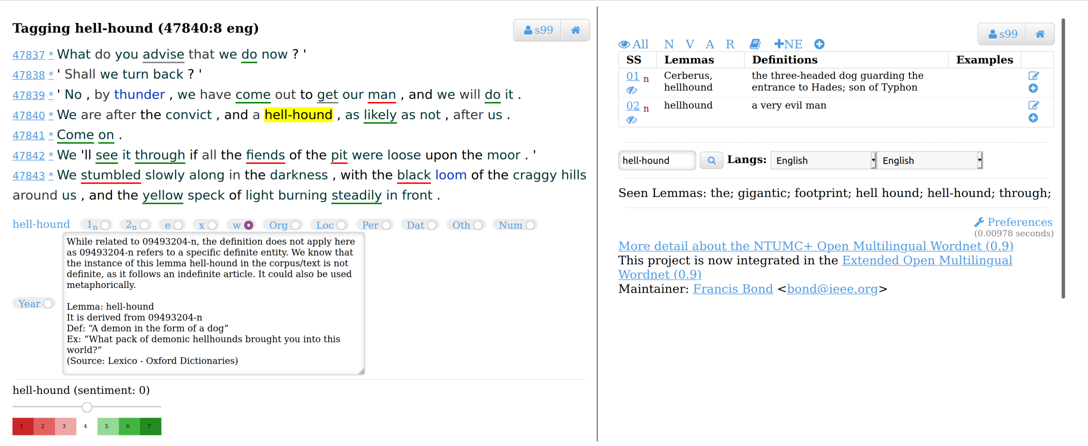
We used an enhanced version of the annotation tool IMI described in Bond et al. (2015). As well as selecting the sense, it allows annotators to tag senses in context with sentiment (from -100 to +100).

Figure 1 shows a passage that has been tagged. The text is shown on the left. Words with positive and negative sentiment are shown with red and green underlines respectively. The annotator thinks that there is no suitable sense for the word being tagged (***hell-hound***) so has suggested a new entry in the comments. Existing senses for ***hell-hound*** are shown on the right.

The students only tag a small sample, so they tag as a **sequential** task: annotating chunks of text word-by-word. **Targeted** tagging (annotating by word type) is known to be more accurate (Langone et al., 2004). Our tool, IMI, supports both and the RAs typically use targeted tagging when they add new senses or correct common errors.

2.4 Wordnets

The senses are tagged with enhanced versions of the Princeton WordNet of English (PWN: Fellbaum, 1998), the Chinese Open Wordnet (COW: Wang



Tagging hell-hound (47840:8 eng)

47837 * What do you advise that we do now ? '
47838 * ' Shall we turn back ? '
47839 * ' No , by thunder , we have come out to get our man , and we will do it .
47840 * We are after the convict , and a hell-hound , as likely as not , after us .
47841 * Come on .
47842 * We 'll see it through if all the fiends of the pit were loose upon the moor . '
47843 * We stumbled slowly along in the darkness , with the black loom of the craggy hills around us , and the yellow speck of light burning steadily in front .

hell-hound

While related to 09493204-n, the definition does not apply here as 09493204-n refers to a specific definite entity. We know that the instance of this lemma hell-hound in the corpus/text is not definite, as it follows an indefinite article. It could also be used metaphorically.

Lemma: hell-hound
It is derived from 09493204-n
Def: "A demon in the form of a dog"
Ex: "What pack of demonic hellhounds brought you into this world?"
(Source: Lexico - Oxford Dictionaries)

hell-hound (sentiment: 0)

SS	Lemmas	Definitions	Examples
01 n	Cerberus, hellhound	the three-headed dog guarding the entrance to Hades; son of Typhon	
02 n	hellhound	a very evil man	

hell-hound **Langs:** English English

Seen Lemmas: the; gigantic; footprint; hell hound; hell-hound; through;

(0.00978 seconds)

More detail about the NTUMC+ Open Multilingual Wordnet (0.9)
This project is now integrated in the Extended Open Multilingual Wordnet (0.9)
Maintainer: Francis Bond <bond@ieee.org>

Figure 1: The Sequential Tagging Interface

and Bond, 2013), the Wordnet Bahasa (Bond et al., 2014) and the Japanese wordnet (Isahara et al., 2008). They included systematic extensions for pronouns, chengyu,[2] exclamatives and classifiers (Seah and Bond, 2014; Ho et al., 2014; Morgado da Costa and Bond, 2016) extended with many new senses and semantic relations. For English, 71% of taggable words are tagged with PWN senses, 23% are pronouns, 3.2% are named entities and 2.5% are other new senses we have added.

3 The Student Experience

In this section we provide a summary of students' feedback. Additionally, two students, one each from the linguistics and general elective classes talk about their experience, both as students and later as research assistants.[3]

3.1 Linguistics Students

The students who attained higher grades overall clearly enjoyed the task more. This was evident in their reading the entirety of the text (rather than only the portions assigned to them), and the time they took to deliberate their chosen tags. Several students reported that reading the whole passage through was very useful in helping them situate words, especially polysemous ones, within the broader textual context. Some found tagging only one meaning to be restrictive when multiple interpretations are possible; this reflects students' sensitiveness to multi-faceted words. The inter-annotator comparison segment was useful in resolving doubts and gaining insights towards fine-grained sense distinctions. Some students drew on their knowledge of other languages in referring to the multi-lingual gloss to distinguish between relatively similar words. Overall, student feedback suggests learning from wordnet tagging was a novel and enjoyable experience. Students' active involvement in research thus seems to benefit the processes of teaching, learning, and research.

Linguistics Student's Personal Experience

In the iteration of HG2002 I (Andrew) participated in, the cohort worked on the English-language version of *The Hound of the Baskervilles* by Sir Arthur Conan Doyle. Each section of the corpus was assigned to a pair of students, who would first tag the section without consulting each other. An automated naive annotator (a computer assigning the most frequent sense tag to each lemma) would also tag that section of the corpus (hereafter MFS). We were then presented with an automatically generated list of lemmas for which at least one of the three of us (two humans and one computer) had selected a tag that didn't match the others' choices, and given the go-ahead to discuss our choices with each other. We then worked to come to a consensus (amongst the two human participants) as to the most appropriate tag in each case.

As linguistics students with strongly held opinions and feelings about how language behaves and what words mean, it was useful to have the naive annotator as a third party. For my human annotation partner and myself, the MFS became a kind of common enemy that could not defend itself, and which could generally be relied on to be a worse tagger than we were. When discussing the points of conflict my an-

[2] 成語 *chengyu* "Chinese four character idioms".

[3] Note that students have the option to opt their data out any time up to one week after they get their results. So far no student has asked to do this.

notation partner and I had over our tagging choices (a process that she at one point described as "arguing"), we could generally at least fall back on agreeing that, whatever it was, the computer's choice was probably wrong.

However, the computer annotator was useful as more than a scapegoat. Its choices often did agree with ours (though we spent much less time discussing those cases, as there was no disagreement), affirming both its competence and our own. It was also most interesting to me when its mistakes exposed its own workings. For example, it failed to recognise *finger-tips* as linked to the lemma ***fingertip***, leading me to realise that while punctuation does not usually alter a word's surface form past recognition for a human reader, it might do so for a computer. It was also interesting to me that, while I assumed that a computer program would abide strictly by procedures, its behaviour flaunted some of the instructions we were given as annotators. For example, we were instructed to only tag the highest level of meaning in a multiword expression, as in ***whip up*** as a single lemma, with ***with*** and ***up*** individually marked as '**x**'. However, the computer annotator would routinely assign meaningful tags to both (or multiple) levels.

As someone who grew up reading and enjoying the Sherlock Holmes stories, I was delighted to hear that we would be using them as the source material for this exercise. I also assumed that I would have no trouble with tagging any of the words in the story, as I did not think the language was particularly challenging or archaic. However, once I began using the wordnet, I realised that my initial assumption was far from correct.Beyond simply being familiar with the connotations and denotations of words, and the ways in which they are used, the exercise demanded that I be able to pick out the precise shades of meaning being invoked in any particular instance. Coming from a background of enjoying both literary analysis and creative writing, in which ambiguous or multiple coexisting meanings are rarely subjected to forcible disambiguation, this was an unexpected paradigm shift for me.

A particularly interesting case in which my ideas about fine-grained meaning were challenged was in tagging the lemma ***unimaginative*** in the context of the phrase *practical and unimaginative*. My annotation partner and I both took the collocation of *practical* (which we agreed indicated an interest in concrete concerns) with *unimaginative* into consideration in choosing a sense for ***unimaginative***. I thought that the collocation meant that the two words should have similar senses (thus interpreting ***unimaginative*** as indicating a concern with concrete facts), as two similar ideas placed together for rhetorical emphasis. However, my annotation partner thought that the collocation meant that the two words should have different senses (thus interpreting ***unimaginative*** as "uncreative"), so as to avoid redundancy. Our disagreement in this instance led me to reflect on the ways I use and interpret language in ways I had not previously considered.

Selecting particular senses was an important part of the annotation process. In pursuing this task, we were also forced to attend to the parts of the corpus which we were **not** meant to annotate, including dummy pronouns, auxiliary and modal verbs, conjunctions, and prepositions. While the documentation we were provided with clearly explained that these items should not receive semantically meaningful tags (and should instead be tagged as '**x**'), we were not always clear about what fell into these categories. While some of this confusion was simply reflective of our inexperience at the time, in many cases we felt that leaving these items without meaningful tags would be omitting important semantic information. This was particularly true of modal verbs and prepositions, as we felt that they contributed significantly to the text's meaning. In the case of prepositions, we also faced some confusion, as some more complex prepositions did appear to be available as tags in the wordnet.

This attention to what should not receive meaningful tags alongside what should also revealed to me how closely interdependent the tags (and by extension, the interpretations) we chose were. For example, in dealing with the phrase *were set forth*, the first word (*were*, for the lemma ***be***) should be tagged as '**x**' as auxiliary verb if ***set forth*** were interpreted as a verbal phrase. However, if ***set forth*** were understood as an adjective, ***were*** would become the main verb and would require an appropriate tag.

Working with the wordnet was ultimately a rewarding experience, both as a way of gaining experience with language in actual use and in terms of feeling like I was able to contribute something to a larger project. I also found the interface enjoyable to use and fun to explore; in many ways, the hyperlinked format reminded me of playing a sort of computer game. Being able to compare my annotation with both a human and non-human partner was also invaluable in terms of prompting me to think more deeply about my strategies in sharing and interpreting meaning.

3.2 General Elective Students

Basing the class on Sherlock Holmes was an attractive factor for the majority of students. Most have previously been acquainted with Holmes through media adaptations, but reading the original stories (a class requirement) was a new experience. They were pleased with Arthur Conan Doyle's usage of innovative phrases such as *swamp adder* and *pea jacket*. It removed the impression of the original Holmes texts as too historically stuffy to be understood in modern times. Additionally, using Holmes as a medium to teach linguistics made the subject's technicalities less daunting for students. A student commented, *"I thought it was a really creative idea and since Sherlock Holmes is really popular, it could easily get students interested in linguistics."* Most students were new to wordnets but were brought up to speed with the clear instructional guide to every assignment.

GE Student's Personal Experience

I (Melissa) recall *Detecting Meaning with Sherlock Holmes* as the most carefree yet meaningful class in my undergraduate studies thus far. As a social science major, class content tends towards pessimism. Sociology's assessment mostly takes the form of essays, hence it was refreshing to be graded in this class through another medium (i.e. Wordnet). The class workload was relatively manageable and I could enjoy learning.

Class content was presented in digestible bites of Powerpoint slides, with the right ratio of semantics to more technical linguistic concepts. It was an enjoyable experience of "detecting meaning" with myself, giving names to semantic phenomena I was previously aware of on an intuitive level, but did not know the proper terminology and definitions, especially for the more formal semantics (quantifiers and logical connectives).

On to "detecting meaning" with Sherlock Holmes! The tagging interface is fairly easy to navigate and get accustomed to for a first-time user. I found it rather delightful to dissect words, to pause and ponder its individual meaning, simultaneously separate from and while within the sentence. The assignments took on a personal activity component, as I read through the list of meanings of each word, I referenced them against my personal vocabulary. When I encountered meanings I was previously unaware of, it enhanced the learning factor and expanded my vocabulary. Conversely, I encountered moments of disorientation when the meaning (and sometimes POS) I had in mind was absent from wordnet. On closer inspection, the meaning was often present but tagged with a different morphological form.

The disjunction in meaning took on another dimension during the group project component. Students were grouped with three other classmates who were assigned the same set of sentences in the individual assignment. The task was to confer and settle on one tagged meaning per word. Retagging as a group was an arduous journey for we had varying understandings of the text. Those who spent marginally less time on the first assignment (did not read HOUND in its entirety), tended to tag words literally and out of context. Doubly adding on to the challenge was: One, our visualisation of the story's events were based on different media adaptations of Sherlock Holmes and preexisting knowledge of the Victorian era, or possibly just based on a figment of imagination. Two, our assigned section was a conversation between Dr. Watson and Stapleton as they witnessed a pony get sucked into the Grimpen Mire. It is an abstract conversation when read separately from the main story. Doyle's anthropomorphism of the mire added on to the confusion of whose body part (pony or the mire) some words were referring to.

I was anticipating putting into practice (tagging) everything I learned in class, to encounter and decipher all the possible word puzzle theories. We were assigned 15 sentences each, the length and the literary challenge of which depended on luck. I was a tad disappointed despite knowing it is not feasible for a text to encompass instances of every semantic device. I was hoping for more tagging practice and the chance to make real changes to the corpus beyond proposing suggestions for new entries (part of the assessment criteria). Semantically close reading a text was a new experience, becoming attuned to the finer grains of a text allowed me to forge a deeper appreciation of the effort authors go through in selecting their words.

3.3 Students and Research Output

Most course iterations reveal students who are both outstanding and interested in continuing to contribute to our research goals – something that has happened with the authors of the shared accounts, above. Admittedly, this happens most often with Linguistics students but has also, on occasion, happened with students enrolled in the General Elective course. These longer-term contributions take one of many forms: i) some join the NTU Computational Linguistics Lab as a student research assistant (RA); ii) some decide to write their Final Year Project (FYP) about a related topic; and iii) a selected few join our lab through a program called URECA (Undergraduate Research Experience on CAmpus), designed to cultivate a research culture among the outstanding undergraduate students.

Over the years, our lab has had dozens of student members that were selected from their contributions to the tagging task described in this paper. Most of these students end up making substantial contributions to research problems that emerge and are defined through multiple layers of quality control of the tagging done by our students (discussed in the next section). Some published research that relied on student contributions include: work on Japanese derivational relations (Bond and Wei, 2019); on pronoun representation for Japanese, Mandarin and English (Seah and Bond, 2014); as well as work on exclamatives and classifiers (Mok et al., 2012; Morgado da Costa and Bond, 2016). Other important contributions that came either in the form of theses or research reports include extensive work cleaning up and expanding the Wordnet Bahasa. The resources have been used by students for sentiment analysis (Le et al., 2016; Bond et al., 2019), cross-lingual sense annotation (Bonansinga and Bond, 2016), multilingual crosswords (Tan, 2012) and more.

4 Quality Control and Expert Tagging

Given that the annotation that happens in our classrooms is done by untrained students from diverse backgrounds and often lacking linguistic intuition, it is not surprising that our corpus needs to go through multiple layers of quality control before being suitable for release.

The large majority of this quality control is done by student RAs. This usually happens in phases, and each phase (or RA) focuses on a particular task. These different tasks include: i) review comments left by students during their tagging exercise (e.g. references to possible metaphors, named entities, etc.); ii) review and fix the corpus where problems concerning lemmatization or corpus structure were flagged (i.e. **e** tags); iii) review and address reported gaps in the wordnet coverage (i.e. **w** tags); iv) ensure students made adequate use of the tag **x** (i.e. using it only for words that should not be tagged); and v) review and retag any mistakes in the student annotations. Much of this work ends up rejecting the suggestions made by students, as they often identify real issues without finding the best solution, due to unfamiliarity with wordnets.

To accomplish these tasks, student RAs make use of a set of tools not usually available to other students, including the **targeted** tagging tools (introduced above); the Corpus Fixer which allows the annotator to change the tokenization, POS and lemmatization, as well as to add new multi word expressions; and OMWEdit which allows the annotator to add to or change the wordnets. (Morgado da Costa and Bond, 2015). Some of the non-intuitive aspects of these tools require some training before they can be used but, most importantly, require a deeper understanding of many layers of lexical analysis (e.g. POS tags, lemmatization, multi-word expressions, etc.).

Student RAs without a computational background are often both baffled and amused with problems caused by POS and lemmatization issues (e.g. when words like *graves* are lemmatized as *graf* through a misapplication of the same rule that produces *shelf* from *shelves*), but are quick to grasp these more mechanical aspects of the quality control process.

Most of the other tasks involve more difficult problems, such as judging whether an expression is compositional or not, or whether a distinction in meaning is significant enough to warrant the creation of a new synset. Wordnets are fairly complex, and our student RAs learn about it *on the job*. The task of changing a wordnet feels quite daunting at first, and it only becomes easier once our RAs get familiarized with the wordnet's structure.

Once the decision to create a new synset is made, other layers of complexity arise. Our RAs have to balance the coverage of new senses (i.e. how broad or narrow should the new synset be – taking into consideration other existing synsets). Finding the appropriate semantic links between new and preexisting synsets is also not always straightforward. If the decision is to try to use an existing synset to accommodate a missing sense, then there are other issues to take into account. The main concern is the extent to which an existing synset can be edited to accommodate this alternative meaning. This often requires detailed lexicographic work, observing examples inside and outside our corpus to determine if the proposed changes are warranted by real data.

Many of the more difficult decisions are discussed within larger lab meetings, where multiple student RAs and senior lab members join in. As it was discussed above, some of the problems encountered by our RAs end up deserving a more in depth treatment or discussion, and are taken up by smaller focused teams within our lab, or as the topic of a project/dissertation.

Every time we teach one of the courses described above, a new set of data requiring quality control is created. From our experience, this amounts to roughly 3-4 weeks full-time work for a trained annotator for 600-700 sentences of text. This is often done taking into consideration the written reports submited by students (when available), which also gives our RAs an insight into the common problems that faced student annotators. Whenever possible, these insights are also used to improve the documentation made available to students during their annotation task – with the goal of making this documentation intuitive for students who may feel overwhelmed by the amount of information they need to absorb.

This is not the most efficient way to annotate text, but a good result is obtained in the end, and we can involve many students. One problem we found was that as we refined the tokenization and wordnet guidelines, the corpora got out of sync. For example, when we added pronouns, we had to go back and tag them in the older corpora. More interestingly, we occasionally change our tokenization guides: *long-legged* we used to tokenize as *long* and *-legged* but now tokenize as *long*, -, *-legged*. We also need a new tag for the noun: NND (noun inflected like a pas-participle), which we lemmatize to ***leg***. We are currently working on further using the tagged corpora to find examples in this class; as a source text in corpus linguistics, and for the digital edition of the tagged stories.

4.1 Multilingual Tagging

Many of our student RAs are confident enough to tag and review tagging in other languages present in our corpus (i.e. Mandarin, Japanese, Indonesian or Malay). When this happens, in addition to the quality control process described above, these students are also paid as expert taggers and tag data using their language of choice.

The corpora are made available at `https://github.com/bond-lab/NTUMC/`..

4.2 Dynamic Resources

Our research on lexical semantics is part of a broader attempt to understand language, where we also look at syntax and lexical semantics. Oepen et al. (2004) show that treebanking is an essential part of grammar development — identifying the correct parses from the grammar for a large corpus is the best way to verify its correctness. They suggest a cyclical model of grammar development, where the grammar is revised based on the results of treebanking and then the treebank is updated with the new grammar. To achieve complete coverage, many iterations are necessary. In the same way, we consider sense tagging the best way to verify the coverage and correctness of a wordnet.

Our tagging process looks something like that shown in Figure 2. (i) First the text is pre processed: tokenized, POS tagged and lemmatized. (ii) Then multiple annotators annotate a passage independently, making notes about issues with the corpus or wordnet. (iii) They then compare their annotations and discuss their differences and possibly write up a report. This is the end of the teaching. (iv) The instructor and some RAs go through all entries with comments or as errors. Where necessary, they fix the corpus and/or the wordnet. (v) Finally (although in practice often simultaneously with the previous step) they retag the corpus with the fixed tokenization and lemmatization using the enhanced wordnet. This is then repeated for the next class. The new students start off with a better wordnet, and potentially better preprocessing, tagging tools and guidelines, as enhancements are made based on last year's issues. Thus their task should be easier and the final annotated text better. This is similar to the **spiral** model of software development described by software developers such as Boehm (1988); Gilb (1989); Larman and Basili (2003). At each loop the development cycle (here we consider we are developing the wordnet, corpus and tools) the process becomes gradually better.

We feel that the wordnets needs to go through several more iterations of tagging and fixing before all the commonly appearing issues are fixed, and of course annotation in new domains will bring new families of problems. One non-trivial problem is coordinating our improvements with others: we are

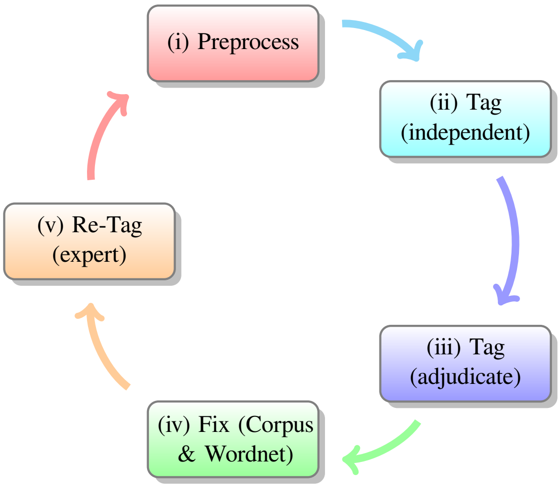


Figure 2: Sense Annotation Spiral

doing our best to coordinate with the English Wordnet (McCrae et al., 2019) and linking through the Collaborative Interlingual Index (CILI Bond et al., 2016b). However, this integration is not seamless.

There are still many questions left unsolved. We still have many lexical semantic phenomena not covered: auxiliary verbs, conjunctions and prepositions; light verb+noun combinations; decomposable semantics (e.g. *unADJ* is productively the antonym of *ADJ*); multiple interpretations, ...These are often taken up by students as final year projects or research projects in other classes.

5 Conclusion and Future Work

We need more annotated text: linking text to analysis is an important task. We expect linking to lead to changes in the linked resource: it is important to support this. Access to more data makes more interesting projects possible. Students learn a lot by attempting real tasks, and enjoy working on interesting stories. We can take advantage of this to improve the quantity and quality of our wordnets and corpora.

One of the goals of this paper is to encourage other similar courses around the world to integrate similar strategies to annotate more text. We have had success supporting colleagues at the University of Pisa in order to tag an Italian translation of *the Speckled Band* as art of a semantics course. We would like to like to coordinate with more lecturers in other countries to extend the task to other languages. This is also why we commit to open-source practices, and make both our data and our tools[4] available on GitHub.

Acknowledgements

This paper started off as an invited talk at the Linked Data in Linguistics workshop (LDL 2018). We would like to thank all the students involved in the tagging: HG2002 (2011–2015, 2018–2020) HG8011 (2016,2018,2019), the URECA projects, FYPs and student assistants. And of course we thank all the wordnet developers, especially Christiane Fellbaum. We would like to thank Alessandro Lenci, Giulia Boanasinga and Tommaso Petrolito for their help with preparing and tagging the Italian data.

For ideas and information about Sherlock Holmes, we would like to thank the Sound of the Baskervilles for their encouragement and advice and Alexis Barquin of the Arthur Conan Doyle Encyclopedia, who kindly allowed us to access his database of texts.

Some of this work was funded by the following grants: MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13, Singapore), *Inter-lingual equivalence in lexical databases* (NSC, Poland) and *Joint Research on Multilingual Semantic Analysis* (Fuji-Xerox Corporation, Japan).

References

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (www.nltk.org/book).

Stephen Bird, Ewan Klein, and Edward Loper. 2010. *Nyumon Shizen Gengo Shori [Introduction to Natural Language Processing]*. O'Reilly. (translated by Hagiwara, Nakamura and Mizuno).

Barry Boehm. 1988. A spiral model of software development and enhancement. *IEEE Computer*, 21(5):61–71.

Giulia Bonansinga and Francis Bond. 2016. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, pages 45–49.

Francis Bond, Arkadiusz Janz, and Maciej Piasecki. 2019. A comparison of sense-level sentiment scores. In *Proceedings of the 11th Global Wordnet Conference (GWC 2019)*.

[4] https://github.com/bond-lab/NTUMC/ (data) https://github.com/bond-lab/IMI/ (tools)

Francis Bond, Lian Tze Lim, Enya Kong Tan, and Hammam Riza. 2014. The combined wordnet Bahasa. *Nusa: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Francis Bond, Luís Morgado da Costa, and Tuấn Anh Lê. 2015. IMI — a multilingual semantic annotation environment. In *ACL-2015 System Demonstrations*.

Francis Bond, Tomoko Ohkuma, Luís Morgado da Costa, Yasuhide Miura, Rachel Chen, Takayuki Kuribayashi, and Wenjie Wang. 2016a. A multilingual sentiment corpus for Chinese, English and Japanese. In *6th Emotion and Sentiment Analysis Workshop (at LREC 2016)*. Portorož.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016b. CILI: The collaborative interlingual index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, pages 50–57.

Francis Bond and Ryan Lim Dao Wei. 2019. Generating derivational relations for the japanese wordnet: The case of agentive nouns. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–7. IEEE.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

Arthur Conan Doyle. 1902. *The Hound of the Baskervilles*. George Newnes, London.

Arthur Conan Doyle. 1905. *The Return of Sherlock Homes*. George Newnes, London. Project Gutenberg www.gutenberg.org/files/108/108-h/108-h.htm.

Arthur Conan Doyle. 1927. How I made my list. *Strand Magazine*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Tom Gilb. 1989. *Principles of Software Engineering Management*. Addison Weslet Longman.

Wan Yu Ho, Christine Kng, Shan Wang, and Francis Bond. 2014. Identifying idioms in Chinese translations. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

David A. Kolb. 1984. *Experiential Learning: Experience as the Source of Learning and Development*. Prentice-Hall.

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating wordnet. In *Workshop On Frontiers In Corpus Annotation*, pages 63–69. ACL, Boston.

Craig Larman and Victor R Basili. 2003. Iterative and incremental development: A brief history. *Computer*, 36(6):47–56.

Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment analysis for low resource languages: A study on informal Indonesian tweets. In *Proceedings of The 12th Workshop on Asian Language Resources*, page 123−131. Osaka.

Lothar Lemnitzer and Claudia Kunze. 2004. Using wordnets in teaching virtual courses of computational linguistics. In *Proceedings of the 2nd Global Wordnet Conference (GWC 2004)*, pages 150–156.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 —an open-source wordnet for English. In *Proceedings of the 11th Global Wordnet Conference (GWC 2019)*.

Hazel Shuwen Mok, Eshley Huini Gao, and Francis Bond. 2012. Generating numeral classifiers in Chinese and Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 211-218.

Luís Morgado da Costa and Francis Bond. 2015. Omwedit - the integrated open multilingual wordnet editing system. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, System Demonstrations (ACL 2015)*, pages 73–78. Beijing, China. URL static/pubs/acl2015-omwedit-demo.pdf.

Luís Morgado da Costa and Francis Bond. 2016. Wow! what a useful extension to wordnet! In *10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož.

Stephan Oepen, Dan Flickinger, and Francis Bond. 2004. Towards holistic grammar engineering and testing — grafting treebank maintenance into the grammar revision cycle. In *Beyond Shallow*

Analyses — Formalisms and Statistical Modelling for Deep Analysis (Workshop at IJCNLP-2004). Hainan Island. URL http://www-tsujii.is.s.u-tokyo.ac.jp/bsa/.

Geoffrey K. Pullum. 1984. If it's tuesday, this must be glossematics. *Natural Language & Linguistic Theory*, 2(1):151–156. URL http://www.jstor.org/stable/4047563.

Eric S. Raymond. 1999. *The Cathedral & the Bazaar*. O'Reilly.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Jeanette Yi Wen Tan. 2012. *Automatic Generation of Multilingual Crossword Puzzles with WordNet*. Final year project, Linguistics and Multilingual Studies, Nanyang Technological University.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

竜之介芥川. 1918. 蜘蛛の糸. 赤い鳥.